Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Аналіз громадської думки методами обробки природної мови (NLP)
(тема)

Виконав:
здобувач _2_ року навчання, групи _САУм-23-2
Зюкін Д.С.
(прізвище, ініціали)
Спеціальність
124 Системний аналіз
(код і повна назва спеціальності)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
Системний аналіз і управління
(повна назва освітньої програми)
Керівник доц. Гибкіна Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Сидоров М.В.
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет _інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _другий (магістерський)
Спеціальність _124 Системний аналіз
(код і повна назва)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Системний аналіз і управління
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
“ _25_ ” _листопада_ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві _Зюкіну Дмитру Сергійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи _Аналіз громадської думки методами обробки природної мови (NLP)

затверджена наказом по університету від 22 листопада 2024 р. № 1228 Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 6 січня 2025 р.

3. Вихідні дані до роботи _набір з трьох датасетів з платформи Kaggle: «US Airline Tweets», «Google Play Reviews» та «Social Media Tweets», що містять текстові дані для аналізу громадської думки з використанням методів обробки природної мови (NLP)

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз предметної області
2. Вибір і обґрунтування методу розв’язання
3. Програмна реалізація
4. Результати обчислювального експерименту
5. Аналіз можливих застосувань

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій

1. Актуальність теми роботи
2. Постановка задачі
3. Системний аналіз предметної області
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	25 листопада – 1 грудня 2024 р.	виконано
2	Вибір та обґрунтування методу	2 – 8 грудня 2024 р.	виконано
3	Розробка алгоритму і програми	9 – 22 грудня 2024 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	23 – 29 грудня 2024 р.	виконано
5	Робота над текстом пояснювальної записки	30 грудня 2024 р. – 9 січня 2025 р.	виконано
6	Представлення роботи на рецензію в ЕК	10 січня 2025 р.	виконано

Дата видачі завдання _25_ листопада 2024 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Гибкіна Н.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 84 с., 14 табл., 21 рис., 1 дод., 9 джерел.

ОБРОБКА ПРИРОДНОЇ МОВИ (NLP), VADER, TEXTBLOB, TRANSFORMERS, SPACY, STANFORD NLP, OPENNLP, АНАЛІЗ ГРОМАДСЬКОЇ ДУМКИ, ВІДГУКИ КОРИСТУВАЧІВ, ТОНАЛЬНІСТЬ ТЕКСТІВ, АВТОМАТИЗАЦІЯ АНАЛІЗУ.

Об’єкт дослідження – процес автоматизованого аналізу громадської думки на основі текстових даних, отриманих із відгуків споживачів.

Предмет дослідження – методи та інструменти обробки природної мови (NLP), що використовуються для аналізу тональності тексту.

Мета роботи – створення програмного інструменту для автоматизованого аналізу громадської думки за допомогою методів обробки природної мови (NLP) та порівняння ефективності різних NLP-бібліотек.

Методи дослідження – методи обробки природної мови для аналізу тональності текстів, системний аналіз для проєктування архітектури програмного забезпечення, методи порівняльного аналізу результатів роботи NLP-бібліотек, методи тестування програмного забезпечення.

У роботі створено програмний інструмент, який дозволяє завантажувати набори даних із текстовими відгуками, виконувати аналіз тональності текстів за допомогою бібліотек NLP (VADER, TextBlob, Transformers, spaCy, Stanford NLP), а також порівнювати отримані результати. Проведено тестування інструменту на наборах даних з платформи Kaggle. Новизна роботи полягає у реалізації підходу до інтегрованого порівняння результатів роботи різних NLP-бібліотек у рамках одного програмного продукту.

Розроблений інструмент може бути використаний для аналізу відгуків у бізнесі, маркетингових дослідженнях, соціології, а також у навчальному процесі для практичного ознайомлення з методами NLP.

ABSTRACT

Introductory note: 84 pages, 14 tables, 21 figures, 1 appendix, 9 sources.

NATURAL LANGUAGE PROCESSING (NLP), VADER, TEXTBLOB, TRANSFORMERS, SPACY, STANFORD NLP, OPENNLP, PUBLIC OPINION ANALYSIS, USER REVIEWS, TEXT SENTIMENT, AUTOMATED ANALYSIS.

Object of the study – the process of automated public opinion analysis based on text data obtained from consumer reviews.

Subject of the study – methods and tools of natural language processing (NLP) used for text sentiment analysis.

Purpose of the study – to develop a software tool for automated public opinion analysis using natural language processing (NLP) methods and comparing the efficiency of various NLP libraries.

Research methods – natural language processing methods for text sentiment analysis, systems analysis for designing the software architecture, methods for comparative analysis of NLP library performance, and software testing methods.

The study developed a software tool that enables users to upload datasets containing textual reviews, perform sentiment analysis of texts using NLP libraries (VADER, TextBlob, Transformers, spaCy, Stanford NLP), and compare the obtained results. The tool was tested on datasets from the Kaggle platform.

The novelty of the work lies in the implementation of an integrated approach to comparing the performance of different NLP libraries within a single software product.

The developed tool can be applied to review analysis in business, marketing research, sociology, and educational processes for practical acquaintance with NLP methods.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАК, ОДИНИЦЬ І ТЕРМІНІВ

NLP – обробка природної мови (Natural Language Processing);

Сентиментальний аналіз – аналіз емоційного тону в текстах;

Бібліотека NLP – програмний інструмент для обробки природної мови;

API – інтерфейс програмування додатків (Application Programming Interface);

DB – база даних (Database);

UI – користувацький інтерфейс (User Interface);

CSV – кома-роздільні значення (Comma-Separated Values);

HTTP – протокол передачі гіпертекста (HyperText Transfer Protocol).

ВСТУП

Актуальність теми. Сучасний розвиток інформаційних технологій, зокрема розширене використання соціальних медіа та цифрових платформ, створює нові можливості для аналізу громадської думки. У зв'язку з великою кількістю текстових даних, що генеруються користувачами у вигляді відгуків про продукти, послуги та компанії, постає необхідність автоматизованого аналізу цих даних. Методи обробки природної мови (NLP) стають важливими інструментами для виявлення основних трендів і настроїв, що дозволяє ухвалювати обґрунтовані стратегічні рішення на основі отриманих результатів.

Мета і завдання кваліфікаційної роботи. Метою цієї роботи є створення програмного інструменту для автоматизованого аналізу громадської думки за допомогою методів NLP та порівняння різних підходів до аналізу текстових даних.

Для досягнення поставленої мети необхідно виконати наступні завдання:

– провести огляд і аналіз сучасного стану задачі автоматизованого аналізу громадської думки з використанням NLP;

– дослідити існуючі методи та бібліотеки для аналізу тональності текстів, такі як VADER, TextBlob, Transformers, spaCy, Stanford NLP та OpenNLP;

– розробити архітектуру програмного інструменту, що включає бекенд для виконання NLP-аналізу та інтерфейс користувача для зручного доступу до результатів;

– реалізувати функціонал порівняння результатів аналізу різними бібліотеками та провести тестування на різних наборах даних.

Об'єктом дослідження є процес аналізу громадської думки на основі текстових даних, отриманих із відгуків споживачів.

Предметом дослідження є методи та інструменти обробки природної мови для аналізу тональності текстів.

Методи дослідження. У роботі використовуються методи обробки природної мови для аналізу тональності текстів, порівняльний аналіз результатів

різних бібліотек, системний аналіз для розробки архітектури програмного продукту та методи тестування програмного забезпечення.

Публікації. Результати, отримані у кваліфікаційній роботі, було представлено на ІІІ Міжнародній молодіжній науково-практичній конференції англійською мовою «Навчання і викладання: у світі після війни» (м. Харків, 08 листопада 2024 р.) [1].

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Системний аналіз задачі аналізу громадської думки

1.1.1 Вербальна модель системи

У сучасному світі аналіз громадської думки набуває особливого значення через масове поширення інтернету та соціальних мереж. Система, що розробляється, має на меті автоматизувати процес аналізу текстових даних, які відображають ставлення людей до певних продуктів, послуг або брендів.

Об’єктом дослідження є висловлювання користувачів, представлені у вигляді текстових наборів даних.

Предметом дослідження є методи та інструменти обробки природної мови (NLP), які використовуються для аналізу тональності та емоційного забарвлення текстів.

Точка зору: дослідник, який прагне розробити ефективний інструмент для аналізу громадської думки.

Призначення системи: надання можливості користувачам здійснювати аналіз текстових даних за допомогою різних NLP-бібліотек, порівнювати результати та отримувати корисні висновки для прийняття рішень.

Мета системи: створити програмне забезпечення для автоматизованого аналізу громадської думки з можливістю порівняння ефективності різних методів NLP.

Класифікація системи:

– за функціональністю: аналітична система, спрямована на обробку та інтерпретацію даних для виявлення закономірностей;

– за масштабом: залежить від обсягів даних та джерел;

– за типом даних: працює з неструктурованими текстовими даними;

– за ступенем інтерактивності: забезпечує активну взаємодію з користувачем через інтуїтивний інтерфейс.

Особливість цієї системи полягає в її можливості об’єднувати різні NLP-бібліотеки та здійснювати порівняльний аналіз отриманих даних. Це відрізняє її від інших схожих рішень і підкреслює її ключові характеристики.

1.1.2 Морфологічний опис системи

Структура та склад системи:

– набори даних на вході: система приймає готові набори даних, надані користувачем, які містять текстову інформацію для аналізу;

– модуль підготовки даних: здійснює очищення та попередню обробку текстів, включаючи видалення зайвих символів, нормалізацію та токенізацію;

– аналітичний модуль: реалізує функціонал для аналізу тональності та емоційного забарвлення текстів за допомогою різних NLP-бібліотек (VADER, TextBlob, Transformers тощо);

– база даних: зберігає результати аналізу та метадані про проведені дослідження;

– модуль візуалізації: відповідає за представлення результатів у зручному для користувача форматі;

– інтерфейс користувача: забезпечує зручний доступ до функцій системи та відображає результати аналізу.

Межі системи та взаємодія із зовнішнім середовищем:

– користувачі: взаємодіють із системою через інтерфейс, надають дані для аналізу та отримують результати;

– зовнішні бібліотеки та сервіси: використання сторонніх NLP-бібліотек вимагає сумісності та належної інтеграції.

Тепер представимо модель системи аналізу громадської думки у вигляді «чорної скриньки». Ця модель дозволяє спростити уявлення про систему, зосередившись на її входах та виходах, не розкриваючи внутрішню структуру та реалізацію.

Вхід: набори даних з текстовою інформацією, надані користувачем. Це можуть бути відгуки, коментарі, пости в соціальних мережах тощо.

Вихід: проаналізовані результати з визначенням тональності (позитивна, негативна, нейтральна) та емоційного забарвлення текстів.

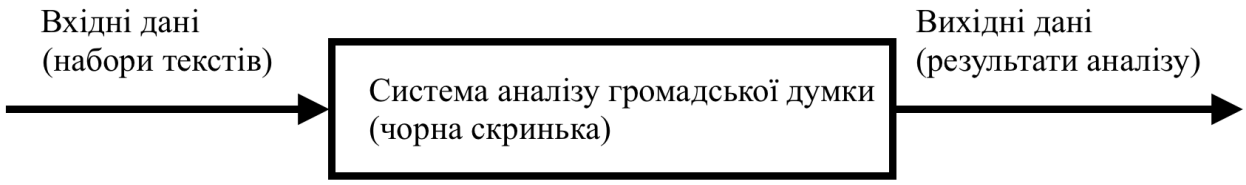

Рисунок 1.1 – Модель типу «чорна скринька»

На рисунку 1.1 показано модель «чорної скриньки» системи аналізу громадської думки. Вхідними даними є набори текстових даних, надані користувачем. Ці дані можуть містити різноманітну інформацію, таку як відгуки про продукти, коментарі на форумах або пости в соціальних мережах.

Система аналізу громадської думки, яка виступає в ролі «чорної скриньки», отримує ці вхідні дані та виконує їх обробку за допомогою методів обробки природної мови. Внутрішня робота системи не деталізується в цій моделі, що відповідає концепції «чорної скриньки». Система застосовує різні NLP-бібліотеки для аналізу тексту, але на цьому етапі ми не зосереджуємося на тому, як саме це відбувається.

На виході користувач отримує проаналізовані дані, що містять інформацію про тональність текстів (позитивну, негативну або нейтральну) та емоційне забарвлення, включаючи виявлення специфічних емоцій, таких як радість, сум, гнів тощо. Ці результати надаються користувачу у зручному форматі.

Таким чином, модель «чорної скриньки» дозволяє сфокусуватися на основних функціях системи, підкреслюючи її взаємодію з користувачем та кінцевий результат, без деталізації внутрішніх механізмів роботи. Це спрощує розуміння призначення системи та її користі для кінцевого користувача.

Для повного розуміння функціонування системи аналізу громадської думки важливо не лише розглянути її внутрішні компоненти, але й зрозуміти, як вона взаємодіє із зовнішнім середовищем. Модель зовнішнього середовища системи допомагає виявити основні зовнішні фактори та елементи, які впливають на роботу системи або з якими вона взаємодіє.

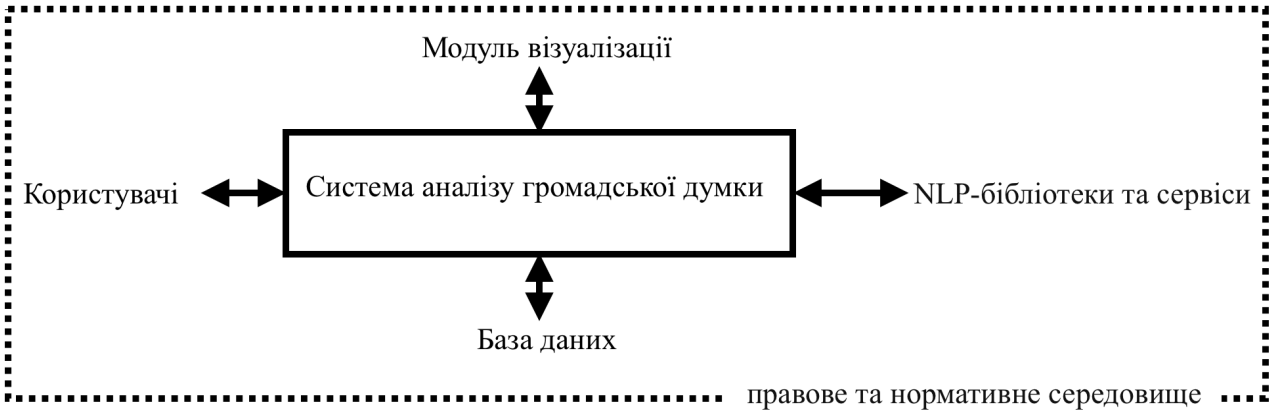

Рисунок 1.2 – Модель зовнішнього середовища системи

На рисунку 1.2 представлена модель зовнішнього середовища системи аналізу громадської думки. У центрі схеми розміщена сама система, яка взаємодіє з різними зовнішніми компонентами.

Основними елементами зовнішнього середовища системи є користувачі, NLP-бібліотеки та сервіси, правове та нормативне середовище, а також база даних результатів. Користувачі взаємодіють із системою через інтерфейс користувача, завантажуючи набори даних для аналізу та отримуючи результати. Вони можуть бути дослідниками, маркетологами, аналітиками або будь-якими іншими зацікавленими особами.

Система використовує зовнішні бібліотеки для обробки природної мови, такі як VADER, TextBlob, Transformers тощо. Ці NLP-бібліотеки та сервіси надають інструменти та алгоритми для аналізу тексту, що дозволяє системі ефективно виконувати свої функції.

Важливим аспектом є відповідність системи вимогам правового та нормативного середовища щодо захисту персональних даних та авторських прав.

Це включає дотримання політик конфіденційності, отримання необхідних згод від користувачів та забезпечення безпеки зберігання і передачі даних.

Крім того, система містить базу даних результатів, яка використовується для збереження проаналізованих даних. Це дозволяє користувачам отримувати доступ до результатів аналізу для подальшого використання, що сприяє більш глибокому розумінню громадської думки та підтримує процес прийняття рішень.

1.1.3 Функціональна модель системи

Функціональна модель системи аналізу громадської думки відображає основні процеси та взаємодії, які забезпечують виконання системою своїх функцій. На рисунку 1.3 представлена функціональна діаграма системи, створена в ARIS.

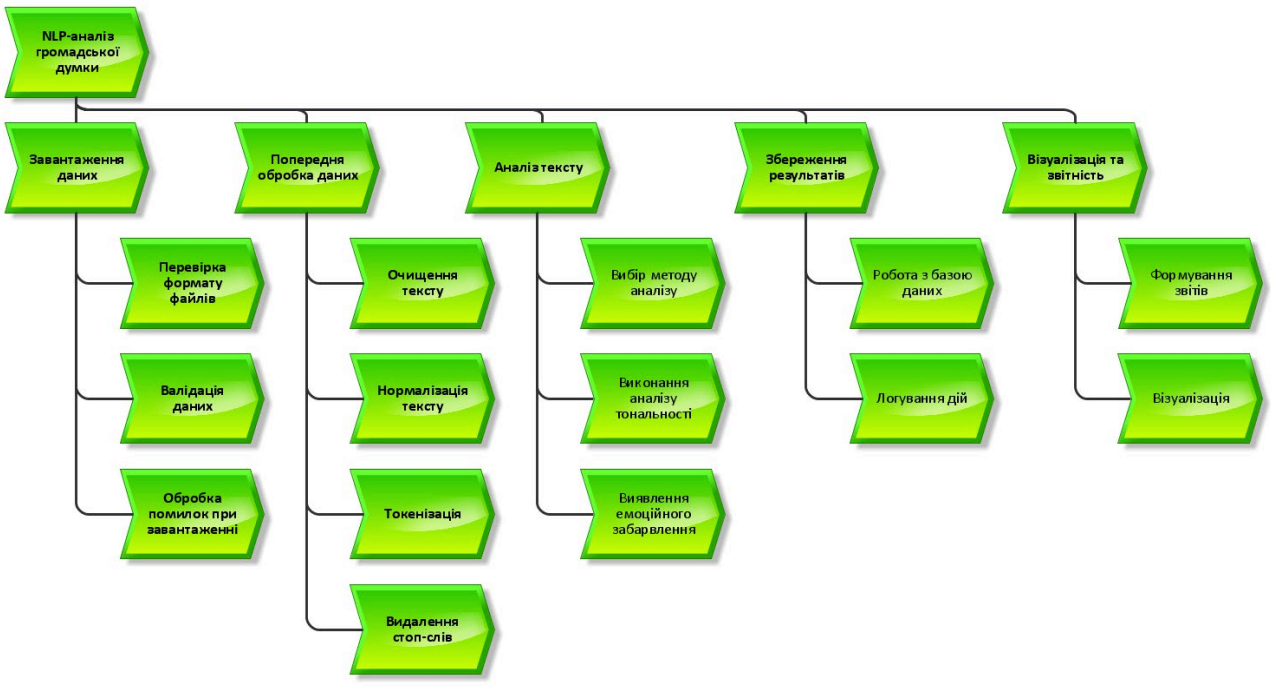

Рисунок 1.3 – Функціональна модель системи у вигляді Process Landscape діаграми в ARIS

Основні процеси системи включають завантаження даних, попередню обробку, аналіз тексту, збереження результатів та візуалізацію з формуванням звітів. Спочатку система здійснює завантаження даних, під час якого перевіряє формат файлів, щоб упевнитися, що вони відповідають підтримуваним форматам. Далі виконується валідація даних, щоб перевірити їх цілісність, правильність і наявність усіх необхідних полів, а також відсутність критичних помилок.

Після завантаження даних здійснюється попередня обробка тексту. На цьому етапі дані очищуються від зайвих символів, HTML-тегів та пунктуації, щоб забезпечити більш точний аналіз. Потім виконується нормалізація тексту, яка включає приведення до нижнього регістру та усунення спеціальних символів. Далі система розбиває текст на окремі слова або токени під час токенізації та видаляє стоп-слова, тобто поширені слова, які не несуть змістового навантаження.

Етап аналізу тексту передбачає вибір відповідного методу або NLP-бібліотеки для проведення аналізу. Система визначає полярність тексту, класифікуючи його як позитивний, негативний або нейтральний, а також виявляє емоційне забарвлення, ідентифікуючи специфічні емоції, такі як радість, сум або гнів.

Після аналізу результати зберігаються в базі даних, що забезпечує можливість подальшого доступу до них. Система також фіксує всі дії користувача та процеси для забезпечення безпеки та можливості аудиту.

На заключному етапі відбувається візуалізація результатів та формування звітів.

Ця модель служить основою для розуміння роботи системи та її подальшого вдосконалення.

1.1.4 Інформаційна модель

Інформаційна модель системи аналізу громадської думки ілюструє структуру даних і взаємозв’язки між основними сутностями, що використовуються

для зберігання та обробки інформації в системі. У процесі роботи з текстовими даними система проходить кілька основних етапів: завантаження даних, вибір методу аналізу, виконання аналізу тексту та створення звітів на основі отриманих результатів. Модель побудована таким чином, щоб забезпечити ефективне управління всіма цими процесами.

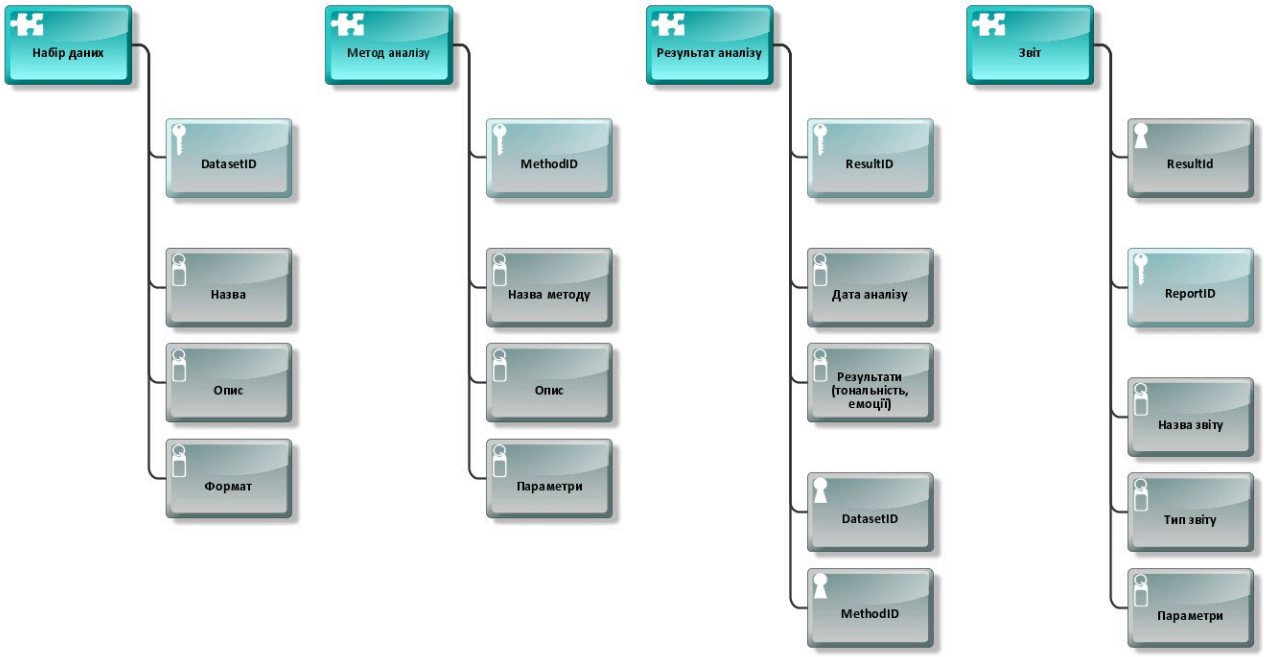

Рисунок 1.4 – Інформаційна модель системи у вигляді
Data model діаграми в ARIS

Сутність «Набір даних» зберігає інформацію про завантажені текстові дані, які будуть аналізуватися. Кожен набір даних має унікальний ідентифікатор (DatasetID) і містить такі атрибути, як назва, опис і формат. Ці дані є вихідною інформацією, з якою працює система. Наступною важливою сутністю є «Метод аналізу», яка описує методи, що використовуються для аналізу текстових даних. Вона включає унікальний ідентифікатор (MethodID) і містить атрибути, такі як назва методу, опис і параметри. Метод аналізу визначає, як оброблятимуться дані, і може включати використання різних NLP-бібліотек, наприклад VADER, TextBlob, Transformers тощо.

Результати аналізу зберігаються в сутності «Результат аналізу», яка має

унікальний ідентифікатор (ResultID) і містить зовнішні ключі, що пов’язують її з набором даних та методом аналізу. Крім того, ця сутність включає дату проведення аналізу та результати, такі як визначена тональність (позитивна, негативна або нейтральна) і виявлене емоційне забарвлення тексту (радість, сум, гнів тощо). На основі цих результатів створюються звіти, які зберігаються в сутності «Звіт». Ця сутність має унікальний ідентифікатор (ReportID) і містить інформацію про назву звіту, тип та параметри. Звіти дозволяють зберігати результати аналізу у зручному вигляді для подальшого використання.

Взаємозв’язки між цими сутностями забезпечують логічний порядок обробки даних у системі. «Набір даних» пов’язаний з «Результатом аналізу» відношенням «один до багатьох», оскільки один набір даних може мати кілька результатів аналізу. «Метод аналізу» пов’язаний з «Результатом аналізу» відношенням «багато до одного», оскільки один метод аналізу може бути застосований до багатьох наборів даних, але кожен результат аналізу пов’язаний лише з одним методом. Нарешті, «Результат аналізу» пов’язаний з «Звіт» відношенням «один до багатьох», оскільки один результат може бути представлений у вигляді кількох звітів або візуалізацій.

Таким чином, інформаційна модель системи забезпечує чітку організацію та управління даними, необхідними для виконання аналізу громадської думки з використанням методів обробки природної мови. Вона дозволяє ефективно зберігати вихідні дані, налаштовувати методи аналізу, фіксувати результати та створювати звіти для зручного представлення інформації. Модель відображає логіку роботи системи, забезпечуючи її гнучкість та функціональність для різноманітних завдань з аналізу тексту.

1.2 Аналіз сценаріїв вирішення задачі аналізу громадської думки

Проаналізуємо критерії, за якими будемо вирішувати поставлену задачу аналізу громадської думки [2]. У якості критеріїв вибору методу аналізу тексту оберемо наступні:

– критерій 1 (К1): точність визначення тональності та емоційного забарвлення тексту. Важливо, щоб обраний алгоритм забезпечував високу точність у визначенні позитивних, негативних та нейтральних тонів, а також міг чітко ідентифікувати емоційні характеристики тексту;

– критерій 2 (К2): швидкість обробки даних. Оскільки в аналізі громадської думки часто використовуються великі обсяги текстової інформації, необхідно оцінити, наскільки швидко обраний метод може обробляти ці масиви даних;

– критерій 3 (К3): стійкість до шуму в даних. У текстах, таких як коментарі або повідомлення у соціальних мережах, можуть бути орфографічні помилки, сленг або зайві символи, тому метод повинен бути стійким до такого шуму;

– критерій 4 (К4): адаптивність до різних тем. Оскільки громадська думка охоплює різноманітні теми і галузі, метод має легко адаптуватися до нових тем або специфічної лексики;

– критерій 5 (К5): ресурсоємність алгоритму. Важливо врахувати, скільки пам’яті та обчислювальних ресурсів потребує метод, особливо якщо система повинна працювати у реальному часі або на обмеженому обладнанні.

Вибірка альтернатив для обрання методу розв’язання задачі аналізу громадської думки буде складатися з наступної множини значень:

– альтернатива 1 (А1): ручний аналіз. Цей метод передбачає виконання аналізу тексту вручну або з використанням простих інструментів, таких як електронні таблиці чи спеціально налаштовані регулярні вирази. Цей аналіз підходить для невеликих проектів або ситуацій, де важливий глибокий контекстуальний аналіз, однак вимагає значних людських ресурсів і часу, що робить його менш придатним для великих обсягів даних;

– альтернатива 2 (А2): статистичний аналіз тексту. Цей метод базується на використанні простих статистичних підходів, таких як аналіз частоти слів, пошук ключових слів та побудова словників для визначення тональності. Такий підхід може бути ефективним для обробки невеликих обсягів тексту або у випадках, коли потрібна швидка обробка даних, але він часто менш точний у порівнянні з NLP;

– альтернатива 3 (А3): методи обробки природної мови (NLP). Ця альтернатива включає використання сучасних NLP-бібліотек, таких як VADER, TextBlob, Transformers, spaCy та інших. Ці методи дозволяють аналізувати тональність та емоційне забарвлення тексту з високою точністю та можуть адаптуватися до різних видів текстів завдяки широкому спектру моделей.

Побудуємо ієрархічну структуру і відобразимо її на рисунку 1.5.

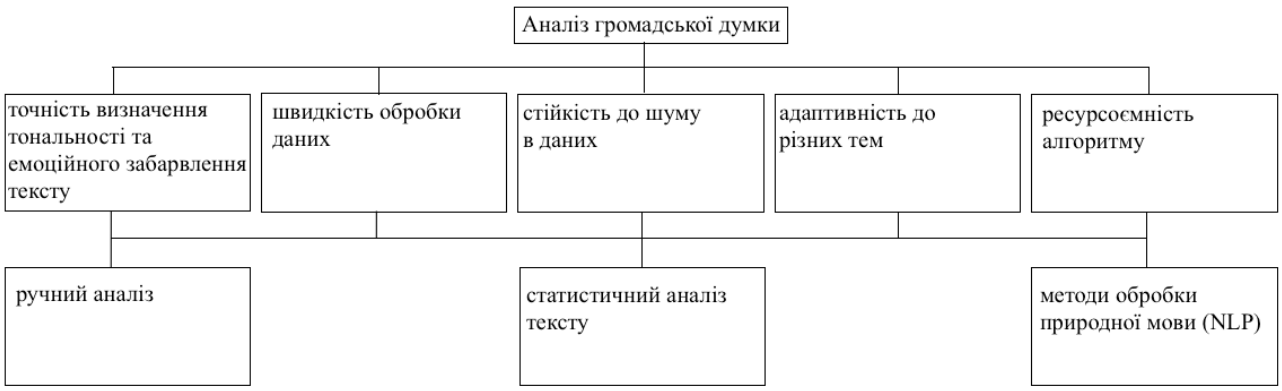

Рисунок 1.5 – Ієрархічна структура

Для аналізу сценаріїв вирішення задачі аналізу громадської думки спершу побудуємо матрицю попарних порівнянь моделі, яка вказана в таблиці 1.1, а також матриці попарних порівнянь критеріїв системи (таблиці 1.2 – 1.6). Ці матриці дозволять оцінити важливість кожного критерію та порівняти альтернативи методів розв’язання задачі між собою. Застосування методу попарних порівнянь забезпечить більш об’єктивний вибір оптимального підходу до аналізу громадської думки, враховуючи вплив різних факторів, таких як точність, швидкість обробки даних, стійкість до шуму, адаптивність та ресурсоємність.

Після побудови матриць ми зможемо визначити ваги кожного критерію, використовуючи метод аналізу ієрархій. Це допоможе прийняти зважене рішення про вибір найбільш ефективного методу, враховуючи всі аспекти аналізу текстів. Далі результати обчислень будуть відображені у вигляді зведених даних, що надасть можливість порівняти ефективність різних підходів і зробити відповідні висновки.

Таблиця 1.1 – Матриця попарних порівнянь критеріїв оцінювання

Критерії оцінювання	К1	К2	К3	К4	К5	Оцінки компонентів	Вектор пріоритетів
К1	1	0,25	0,25	0,2	0,5	0,36	0,06
К2	4	1	2	0,5	4	1,74	0,28
К3	4	0,5	1	0,25	4	1,15	0,18
К4	5	2	4	1	2	2,4	0,39
К5	2	0,25	0,25	0,5	1	0,57	0,09
Всього						6,23	

По даним таблиці 1.1:

– індекс узгодженості

$$IУ = \frac{5,437 - 5}{5 - 1} = 0,109;$$

– відносна узгодженість

$$ВУ = \frac{0,109}{1,12} = 0,1.$$

Таблиця 1.2 – Матриця попарних порівнянь за першим критерієм

К1	А1	А2	А3	Оцінки компонентів	Вектор пріоритетів
А1	1	0,25	0,2	0,37	0,10
А2	4	1	0,5	1,26	0,33
А3	5	2	1	2,15	0,57
Усього				3,78	

По даним таблиці 1.2:

– індекс узгодженості

$$IУ = \frac{3,025 - 3}{3 - 1} \approx 0,012;$$

– відносна узгодженість

$$ВУ = \frac{0,012}{0,58} \approx 0,021.$$

Таблиця 1.3 – Матриця попарних порівнянь за другим критерієм

К1	А1	А2	А3	Оцінки компонентів	Вектор пріоритетів
А1	1	0,2	0,2	0,34	0,09
А2	5	1	0,33	1,18	0,3
А3	5	3	1	2,47	0,62
Усього				3,99	

По даним таблиці 1.3:

– індекс узгодженості

$$ІУ = \frac{3,132 - 3}{3 - 1} \approx 0,066;$$

– відносна узгодженість

$$ВУ = \frac{0,066}{0,58} = 0,114.$$

Таблиця 1.4 – Матриця попарних порівнянь за третім критерієм

К1	А1	А2	А3	Оцінки компонентів	Вектор пріоритетів
А1	1	0,25	0,2	0,37	0,09
А2	4	1	0,2	0,93	0,22
А3	5	5	1	2,92	0,69
Усього				4,22	

По даним таблиці 1.4:

– індекс узгодженості

$$ІУ = \frac{3,217 - 3}{3 - 1} \approx 0,109;$$

– відносна узгодженість

$$ВУ = \frac{0,109}{0,58} \approx 0,187.$$

Таблиця 1.5 – Матриця попарних порівнянь за четвертим критерієм

К1	А1	А2	А3	Оцінки компонентів	Вектор пріоритетів
А1	1	0,5	0,2	0,46	0,12
А2	2	1	0,25	0,79	0,2
А3	5	4	1	2,71	0,68
Усього				3,97	

По даним таблиці 1.5:

– індекс узгодженості

$$IУ = \frac{3,025 - 3}{3 - 1} \approx 0,012;$$

– відносна узгодженість

$$ВУ = \frac{0,012}{0,58} \approx 0,021.$$

Таблиця 1.6 – Матриця попарних порівнянь за п’ятим критерієм

К1	А1	А2	А3	Оцінки компонентів	Вектор пріоритетів
А1	1	0,5	0,25	0,5	0,13
А2	2	1	0,25	0,79	0,21
А3	4	4	1	2,52	0,66
Усього				3,81	

По даним таблиці 1.6:

– індекс узгодженості

$$ІУ = \frac{3,054 - 3}{3 - 1} \approx 0,027;$$

– відносна узгодженість

$$ВУ = \frac{0,027}{0,58} \approx 0,046.$$

У підсумку, аналізуючи отримані значення вектору глобальних пріоритетів (таблиця 1.7), можна зробити висновок, що за умов заданого набору критеріїв для задачі аналізу громадської думки доцільно обрати серед наявних альтернатив третю, а саме метод обробки природної мови (NLP).

Таблиця 1.7 – Глобальні пріоритети

Критерій /Альтернатива	К1	К2	К3	К4	К5	Глобальні пріоритети
А1	0,1	0,09	0,09	0,12	0,13	0,1
А2	0,33	0,3	0,22	0,2	0,21	0,24
А3	0,57	0,62	0,69	0,68	0,66	0,66

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

Аналіз громадської думки є важливим завданням у сучасному суспільстві, особливо в умовах стрімкого розвитку інформаційних технологій та широкого використання соціальних медіа. Велика кількість користувачів залишає свої думки, коментарі та відгуки на різних платформах, створюючи великий обсяг текстових даних, які можуть містити цінну інформацію про настрої та переваги

аудиторії. Завдання полягає у тому, щоб ефективно обробити ці дані, проаналізувати їх тональність (позитивну, негативну чи нейтральну) та виявити емоційне забарвлення, що дозволить краще зрозуміти громадську думку та приймати зважені рішення.

Змістовна постановка задачі передбачає розробку системи, яка здатна обробляти великі обсяги текстових даних, витягувати з них значущу інформацію та аналізувати настрої за допомогою сучасних методів обробки природної мови (NLP). Крім того, одним із важливих аспектів дослідження є порівняння різних NLP-бібліотек, таких як VADER, TextBlob, Transformers, spaCy тощо, для визначення, яка з них забезпечує найкращі результати у контексті точності, швидкості обробки та стійкості до шуму в даних. Основна мета полягає в тому, щоб визначити тональність текстів, виявити емоційні характеристики та порівняти ефективність бібліотек, що дозволить вибрати найбільш оптимальний підхід для різних сценаріїв.

1.3.2 Формальна постановка задачі

Формальна постановка задачі полягає у створенні алгоритмічної моделі, що реалізує аналіз громадської думки на основі вхідних текстових даних D, де D – це множина текстів різних обсягів та формату. Метою є обчислення функції тональності $T(d)$ для кожного тексту $d \in D$, де $T(d)$ набуває значення з множини {позитивна, негативна, нейтральна}, та функції емоційного забарвлення $E(d)$, яка визначає специфічні емоції, такі як радість, сум, гнів тощо. Модель має забезпечувати оптимальний баланс між точністю результатів і швидкістю обробки, використовуючи доступні обчислювальні ресурси.

1.4 Постановка задач дослідження

Змістовна та формальна постановка задачі визначають цілі та ключові вимоги для розробки системи аналізу громадської думки, яка базується на сучасних методах NLP, забезпечуючи ефективну обробку і корисну аналітику для подальшого використання в різних сферах, таких як маркетинг, соціальні дослідження або управління репутацією.

Для успішного виконання поставлених цілей і забезпечення комплексного аналізу громадської думки із застосуванням методів обробки природної мови (NLP) у межах даного дослідження необхідно вирішити кілька ключових задач:

– провести огляд і аналіз сучасного стану задачі автоматизованого аналізу громадської думки з використанням NLP;

– дослідити існуючі методи та бібліотеки для аналізу тональності текстів, такі як VADER, TextBlob, Transformers, spaCy, Stanford NLP та OpenNLP;

– розробити архітектуру програмного інструменту, що включає бекенд для виконання NLP-аналізу та інтерфейс користувача для зручного доступу до результатів;

– реалізувати функціонал порівняння результатів аналізу різними бібліотеками та провести тестування на різних наборах даних.

Виконання цих задач дозволить створити ефективну систему для аналізу громадської думки, а також порівняти різні інструменти NLP, щоб визначити найкращі підходи для аналізу текстових даних.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

Аналіз громадської думки є складним завданням, що потребує ефективних підходів до обробки текстів, які забезпечують високу точність і швидкість аналізу. У цьому розділі представлено основні методи і бібліотеки для автоматизованої обробки текстових даних, що використовуються у сучасних системах NLP для аналізу тональності та емоційного забарвлення тексту. Огляд доступних інструментів дозволяє обґрунтувати вибір компонентів для створення системи автоматизованого аналізу громадської думки.

2.1 Огляд NLP-бібліотек для аналізу громадської думки

Обробка тексту в NLP-системах вимагає вибору бібліотек, що забезпечують різні функціональні можливості для глибокого та багатогранного аналізу текстів. Серед них VADER, розроблений для визначення тональності коротких текстів, таких як коментарі у соціальних мережах, показує себе ефективним інструментом для швидкої оцінки настрою. VADER аналізує позитивні, негативні та нейтральні категорії текстів, забезпечуючи високий рівень швидкості обробки без значного навантаження на систему, що робить його придатним для початкового аналізу [3 – 5].

Інша бібліотека TextBlob надає користувачам простий інтерфейс для роботи з текстами в Python. TextBlob підтримує основні функції лінгвістичного аналізу, зокрема, аналіз частоти слів, визначення тональності, а також базові засоби синтаксичного розбору. Вона підходить для попередньої обробки текстів, особливо в завданнях, де швидкість аналізу є менш критичною.

Бібліотека Transformers від Hugging Face дозволяють використовувати сучасні моделі глибокого навчання, такі як BERT, GPT та інші архітектури, для глибокого контекстуального аналізу тексту. Завдяки цьому підходу Transformers забезпечує високу точність аналізу, дозволяючи моделі не тільки

визначати загальну тональність тексту, а й враховувати деталі семантики та структури, що особливо корисно для довгих текстів з великою кількістю емоційних відтінків.

Бібліотека spaCy є високопродуктивною системою обробки природної мови, яка надає можливості для лінгвістичного аналізу. Вона включає інструменти для розбору частин мови, розпізнавання іменованих сутностей і векторизації слів. Завдяки високій продуктивності spaCy є ефективним вибором для обробки великих обсягів текстів і підходить для професійних додатків, що потребують швидкості обробки.

Stanford NLP пропонує потужний набір інструментів для виконання синтаксичного і семантичного аналізу, що робить його підходящим для наукових досліджень і завдань, де потрібні детальні лінгвістичні знання. Ця бібліотека підходить для задач, які потребують точного розбору складних текстів і вимагають підтримки глибоких мовних моделей.

OpenNLP від Apache є відомою бібліотекою, що забезпечує базовий функціонал для обробки природної мови. Вона підтримує токенізацію, розпізнавання частин мови, розпізнавання сутностей та інші базові NLP-функції, що робить її придатною для завдань загального характеру та попередньої обробки тексту.

2.2 Вибір архітектури та інструментів для реалізації системи

Розробка системи аналізу громадської думки потребує вибору підходу до представлення тексту, що допоможе ефективно обробляти його для подальшого аналізу. Обрана архітектура використовує Python для інтеграції кількох основних методів векторизації тексту, які підходять для реалізації в бібліотеках, таких як VADER, TextBlob, Transformers, spaCy, Stanford NLP та OpenNLP. Кожна з цих бібліотек має свої методи роботи з текстом, що дозволяють застосувати різні підходи до обробки, такі як Bag of Words (BoW), TF-IDF, Word2Vec, GloVe, та трансформерні моделі, включно з BERT [3 – 5].

Bag of Words (BoW) є простим підходом, де текст представляється як набір слів, без врахування порядку. У бібліотеках, таких як TextBlob і OpenNLP, цей метод використовується для базової класифікації та порівняння частотності термінів. BoW дозволяє системі оцінювати, наскільки часто зустрічаються певні слова у текстах, але не враховує контекстуальні особливості, що може обмежувати його точність. Цей підхід підходить для початкової обробки тексту, де потрібно виявити базові тематичні зв'язки, не витрачаючи багато обчислювальних ресурсів.

TF-IDF (Term Frequency-Inverse Document Frequency) вдосконалює BoW, надаючи вищу вагу рідкісним словам, що робить його ефективнішим для виявлення ключових слів і тем. У spaCy та Stanford NLP TF-IDF використовується як метод обробки тексту, що допомагає зосередитися на унікальних термінах, важливих для певних документів. Наприклад, у довгих текстах із домінуванням загальних слів TF-IDF знижує вагу таких термінів, як «і», «в», «з», що допомагає аналізу зосередитися на специфічних поняттях, які визначають зміст тексту.

Word2Vec і GloVe підходять для побудови векторних подань слів, зберігаючи семантичні зв'язки між ними. Ці методи використовуються у бібліотеках spaCy та Stanford NLP, які можуть ефективно будувати зв'язки між схожими словами в багатовимірному просторі. Наприклад, слова «король» і «королева» будуть близькими у векторному просторі, відображаючи їхній семантичний зв'язок. Такі векторні подання дозволяють системі виконувати більш комплексний аналіз ідеї та змісту тексту, розпізнаючи не тільки окремі слова, але й їхні значення у загальному контексті.

Трансформерні моделі, такі як BERT, доступні в бібліотеці Transformers, використовуються для створення контекстних подань тексту. Ці моделі враховують як послідовність, так і зміст тексту, що дозволяє точніше передати значення кожного слова у загальному контексті. Вони особливо ефективні для роботи з довгими текстами, де важливо зрозуміти глибші смисли і деталі. Використання трансформерів в аналізі тексту допомагає врахувати всі значущі елементи, оскільки моделі враховують як попередні, так і наступні слова у реченні.

Побудована система включає інтеграцію бази даних PostgreSQL для зберігання результатів аналізу та забезпечення надійного управління даними. Це дозволяє зберігати результати, генеровані системою, що важливо для подальшого доступу до проаналізованих текстів і результатів.

Інтерфейс користувача (UI), створений на основі React, дозволяє користувачам взаємодіяти з системою, завантажувати дані для аналізу, переглядати результати та порівнювати результати між різними методами.

2.3 Обґрунтування вибору методів для порівняння результатів аналізу

Для вибору оптимального підходу до аналізу громадської думки важливими є критерії точності, швидкості обробки, адаптивності, стійкості до шуму, а також ресурсоємності. Точність аналізу є одним із головних критеріїв, що дозволяє обрати метод, здатний забезпечити високу якість розпізнавання тональності текстів та виявлення емоційних характеристик. Трансформерні моделі зазвичай демонструють високу точність завдяки своїй здатності до контекстуального аналізу, але потребують більше обчислювальних ресурсів.

Швидкість обробки стає важливим фактором при роботі з великими обсягами даних, такими як повідомлення в соціальних мережах. Для таких випадків VADER і TextBlob забезпечують швидке виконання завдань з меншим навантаженням на систему, тоді як трансформери потребують більше ресурсів, але забезпечують точніші результати.

Адаптивність до різних тем і контекстів також є важливим критерієм для аналізу громадської думки, оскільки тексти можуть охоплювати різні галузі. Моделі на основі Word2Vec і GloVe демонструють високу адаптивність, забезпечуючи здатність до розпізнавання зв’язків між словами і концепціями.

Стійкість до шуму є необхідною характеристикою для роботи з текстами соціальних медіа, які можуть містити орфографічні помилки, зайві символи або сленг. Бібліотеки spaCy і Stanford NLP забезпечують додаткові можливості для

фільтрації і коригування таких текстів.

Ресурсоємність є важливим фактором, особливо при обмежених обчислювальних потужностях або необхідності виконувати аналіз у реальному часі. Трансформерні моделі, такі як BERT або GPT, мають високі вимоги до пам'яті та процесора, оскільки для їхнього функціонування потрібні потужні обчислювальні ресурси. Натомість, VADER і TextBlob є менш ресурсоємними, тому підходять для швидкого аналізу, де ресурси обмежені.

Таким чином, поєднання цих бібліотек дозволяє створити комплексну систему для аналізу громадської думки. Бібліотека VADER підходить для швидкого аналізу коротких текстів, TextBlob надає зручний інтерфейс для базового аналізу, Transformers забезпечують глибокий аналіз за допомогою сучасних нейронних моделей, а spaCy пропонує ефективні інструменти для роботи з великими обсягами тексту. Stanford NLP додає можливість точного синтаксичного і семантичного аналізу, тоді як OpenNLP підтримує базові функції NLP.

Використання цих інструментів забезпечує систему аналізу громадської думки з гнучкими можливостями для різних сценаріїв, забезпечуючи надійну обробку текстів із високою швидкістю, точністю та оптимальним використанням ресурсів.

Висновки за розділом 2

У цьому розділі розглянуто основні методи і бібліотеки для аналізу громадської думки, що дозволяють автоматизувати обробку текстових даних з високою швидкістю та точністю. Основними інструментами для цього є VADER, TextBlob, Transformers, spaCy, Stanford NLP та OpenNLP. Кожна з цих бібліотек має свої переваги та обмеження, що робить їх придатними для різних завдань аналізу.

VADER підходить для швидкого аналізу коротких текстів, таких як коментарі в соціальних мережах, забезпечуючи високу швидкість обробки та ефек-

тивність у визначенні тональності тексту. TextBlob пропонує простий інтерфейс для базового лінгвістичного аналізу і є хорошим вибором для попередньої обробки текстів. Transformers дозволяють проводити глибокий контекстуальний аналіз за допомогою моделей глибокого навчання, таких як BERT і GPT, що забезпечує високу точність, але вимагає більших обчислювальних ресурсів.

Stanford NLP і spaCy є потужними бібліотеками для лінгвістичного аналізу і обробки великих обсягів текстів, пропонуючи інструменти для розбору частин мови, розпізнавання іменованих сутностей і семантичного аналізу. OpenNLP забезпечує базові функції для попередньої обробки тексту, такі як токенізація і розпізнавання сутностей, що робить його корисним для загальних завдань.

Комбінування цих бібліотек та методів дозволяє розробити систему аналізу громадської думки, яка буде ефективною, гнучкою та зручною для обробки різноманітних текстів, забезпечуючи точні результати при оптимальному використанні обчислювальних ресурсів.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Загальна характеристика системи

Метою програмної реалізації є створення системи для аналізу громадської думки із застосуванням методів обробки природної мови (NLP). У рамках цього проєкту було розроблено три модулі: бекенд модуль Spring Boot Application, UI модуль React Application, бекенд модуль Python Application, а також була використана база даних PostgreSQL для зберігання наборів даних та результатів аналізу. Ця архітектура забезпечує взаємодію між користувачем, системою та інструментами аналізу NLP.

Spring Boot модуль відповідає за керування бізнес-логікою, обробку запитів користувача та інтеграцію з базою даних. Python модуль реалізує аналітичну частину проєкту, виконуючи аналіз текстів за допомогою різних NLP-бібліотек. Окрім цього, Python модуль активно взаємодіє з базою даних: він зберігає результати обробки текстів та отримує необхідні дані для подальшого аналізу. React модуль забезпечує зручний інтерфейс користувача, дозволяючи обирати датасети, методи аналізу та візуалізувати результати.

Така багаторівнева архітектура гарантує модульність і масштабованість системи. Вона дозволяє легко інтегрувати нові інструменти аналізу, автоматизувати обробку великих обсягів текстів і забезпечувати збереження та доступ до даних у будь-який момент.

3.2 Використані дані

Для аналізу було обрано три датасети з платформи Kaggle, що містять текстові дані з позначенням тональності (позитивна, негативна, нейтральна):

– US Airline Tweets: містить текстові відгуки пасажирів на англійській мові [6];

– Google Play Reviews: включає огляди користувачів мобільних додатків

на англійській мові [7];

– Social Media Tweets: збірка твітів із відповідними оцінками тональності на англійській мові [8].

У таблиці 3.1 наведена загальна інформація про використані датасети.

Таблиця 3.1 – Характеристики датасетів

Датасет	Розмір даних, МБ	Кількість записів	Джерело
US Airline Tweets	1,6	14,845	Kaggle
Google Play Reviews	8,6	18,756	
Social Media Tweets	3,6	25,000	

Ці датасети містять попередньо визначені оцінки, що дозволяє порівнювати результати аналізу NLP-бібліотек із реальними значеннями. Важливо зазначити, що тексти в цих датасетах є різноманітними: одні складаються з одного речення, інші можуть складатися і з кількох. Це додає складності для аналізу, оскільки бібліотеки повинні працювати з різними типами структури тексту. Крім того, у текстах трапляються сарказм, іронія, сленг та складні мовні конструкції, що ще більше ускладнює завдання визначення тональності.

3.3 Архітектура системи

Архітектура системи включає три основні модулі (рисунок 3.1).

Spring Boot Application забезпечує бекенд функціональність для інших модулів системи, реалізуючи управління базою даних, включаючи імпорт, видалення датасетів і запуск процесу аналізу NLP, а також передає метадані аналізу, такі як точність та час виконання, до UI.

React Application виступає користувацьким інтерфейсом, який дозволяє користувачам керувати датасетами, запускати аналіз та переглядати результати, викори-

стовуючи бібліотеку Chart для візуалізації результатів у вигляді графіків і діаграм.

Python Application обробляє запити на аналіз даних з React Application, використовуючи NLP-бібліотеки, такі як VADER, TextBlob, Transformers, spaCy та Stanford NLP, виконуючи аналіз датасетів і зберігаючи результати та час виконання в базу даних для подальшого використання.

PostgreSQL використовується для зберігання наборів даних та результатів аналізу, забезпечуючи надійне збереження, доступ і оновлення інформації для всіх модулів системи.

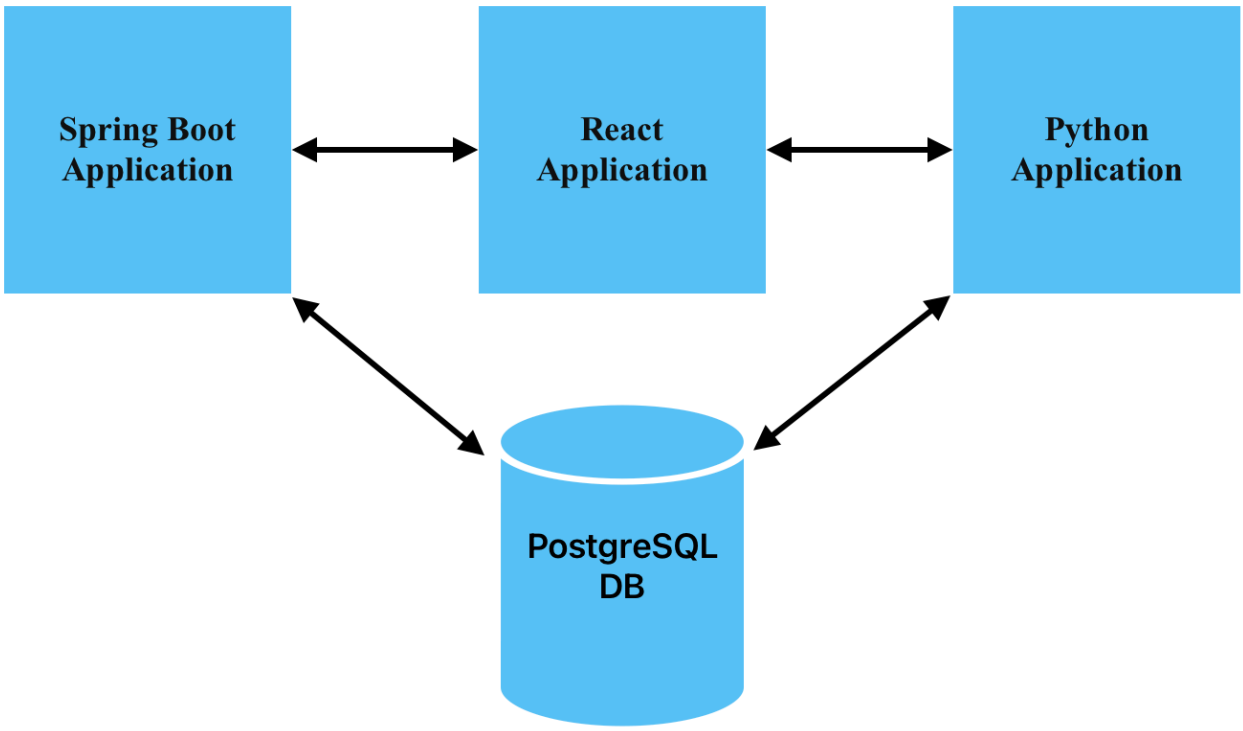

Рисунок 3.1 – Блок-схема архітектури системи

3.4 Реалізація програмних модулів

3.4.1 Spring Boot Application

Spring Boot забезпечує наступний функціонал:

– імпорт датасетів у базу даних (CSV формат);

– видалення датасетів з бази даних;

– забезпечення відображення метрик, тобто побудованих результатів, у вигляді графіків та кругових діаграм.

Ендпоінти Spring Boot API:

– /importDataset – завантаження датасету у базу даних;

– /deleteDataset – видалення датасету з бази даних;

– /getAccuracyMetrics – отримання результатів аналізу точності для побудови графіків;

– /getTimingMetrics – отримання результатів аналізу часу виконання для побудови графіків.

3.4.2 React Application

React Application виконує наступні завдання:

– забезпечує інтерфейс для імпорту, видалення, аналізу датасетів та відображення результатів у вигляді графіків;

– має три основні вкладки: Import / Delete Datasets (рисунок 3.2) – для завантаження, видалення, перегляд списку датасетів; Analyze with NLP (рисунок 3.3) – для запуску аналізу через форму вибору датасетів і бібліотек; Metrics (рисунок 3.4) – для побудови графіків на основі метаданих (для побудови графіків використано бібліотеку Chart.js).

Графіки, які відображаються:

– точність NLP-бібліотек для кожного датасету;

– середня точність NLP-бібліотек по всім датасетам;

– час виконання аналізу для кожної бібліотеки по кожному датасету;

– загальній час виконання кожної бібліотеки по всім датасетам;

– ефективність бібліотек (точність/час) по кожному датасету;

– загальну ефективність бібліотек (точність/час) по всім датасетам.

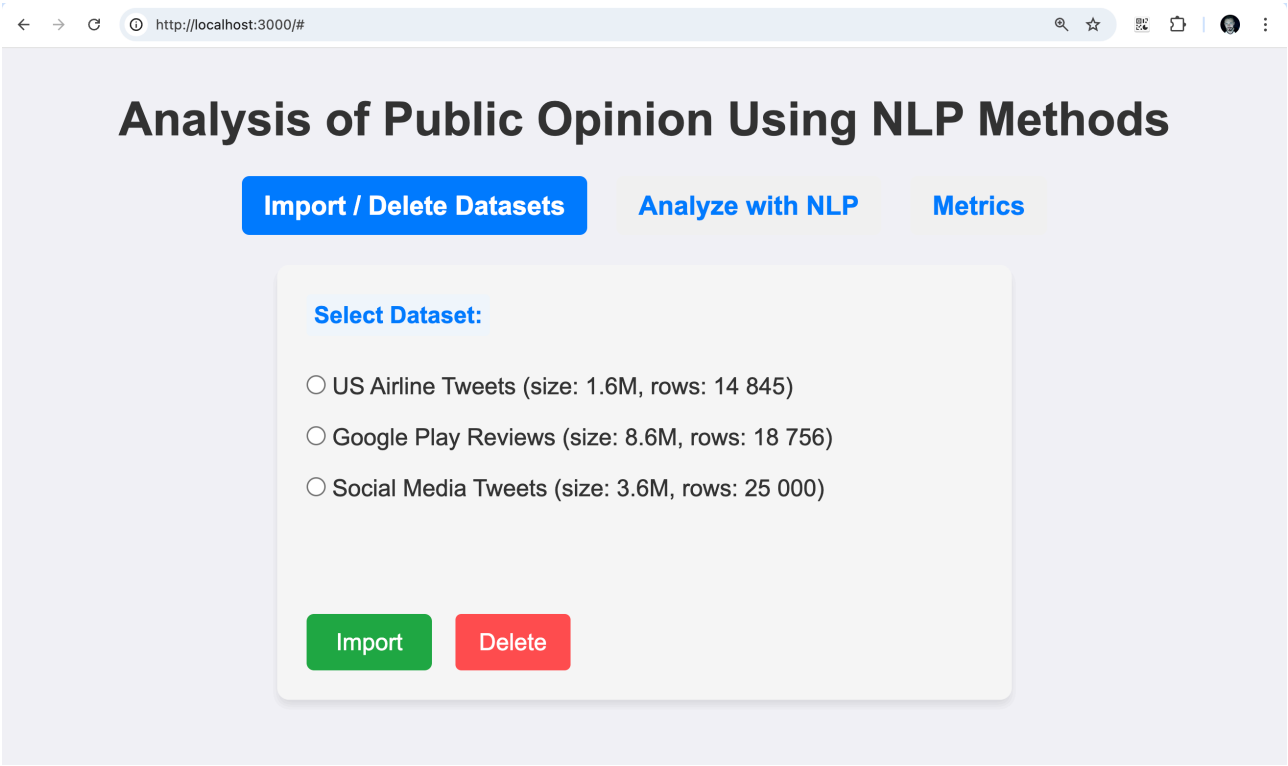

Рисунок 3.2 – Приклад функціональної UI частини, що відповідає за завантаження та видалених датасетів

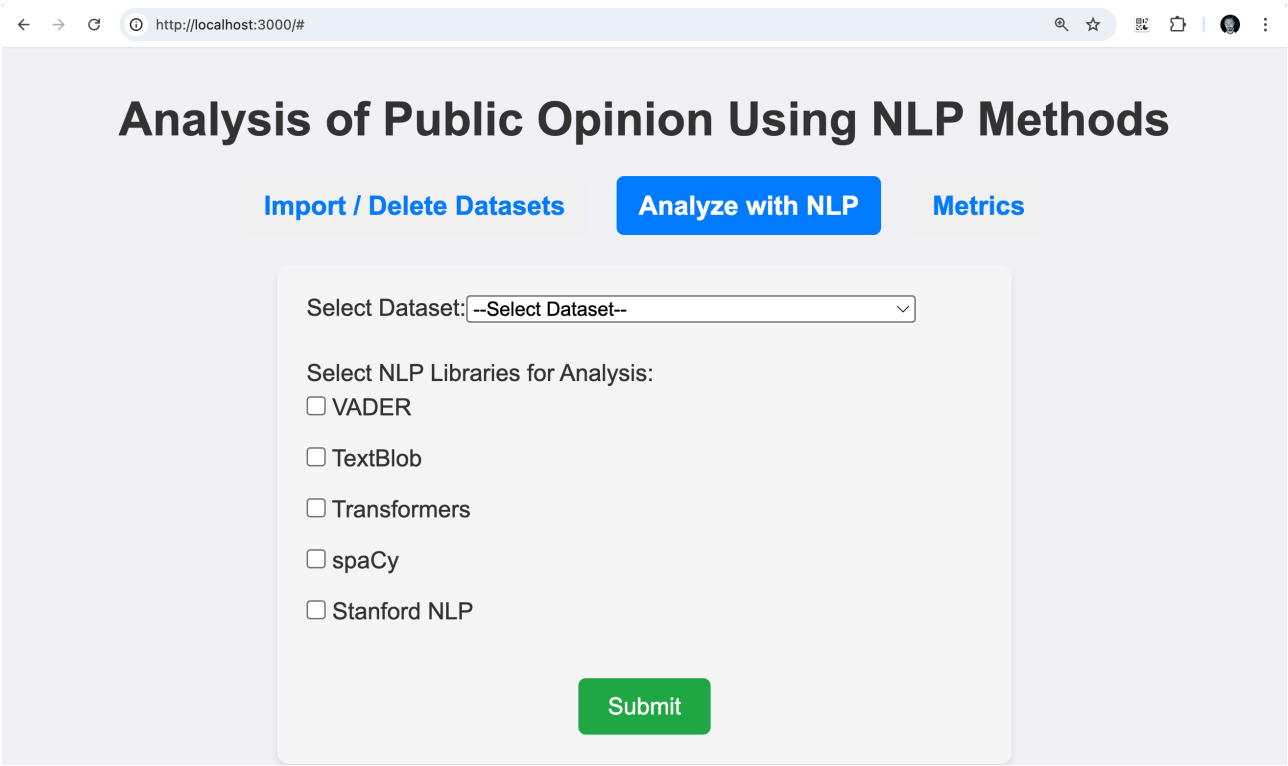

Рисунок 3.3 – Приклад функціональної UI частини, що відповідає за запуск аналізу обраних датасетів обраною NLP-бібліотекою

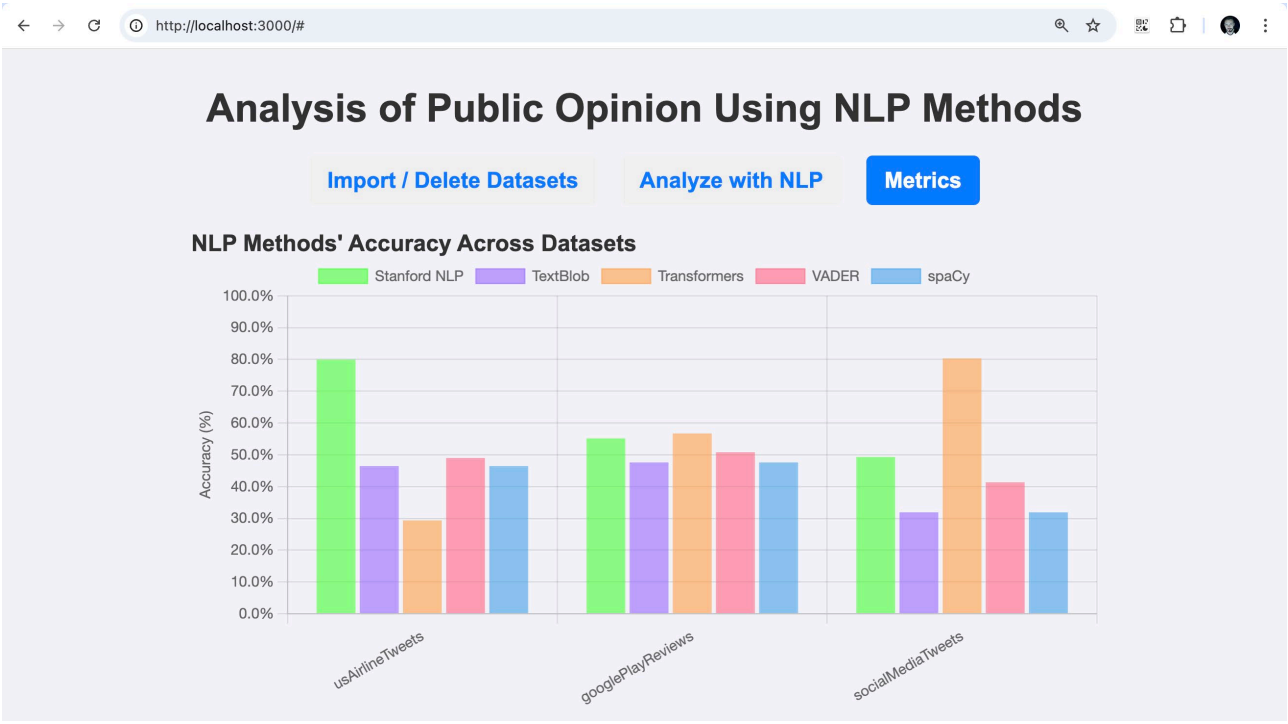

Рисунок 3.4 – Приклад функціональної UI частини, що відповідає за відображення результатів аналізу

3.4.3 Python Application

Python Application реалізує обробку даних за допомогою бібліотек NLP:

– отримує HTTP-запити від React Application;

– витягує дані із бази даних;

– виконує аналіз текстів із використанням бібліотек (VADER, TextBlob, spaCy, Transformers, Stanford NLP);

– зберігає результати аналізу (точність, час виконання) у базу даних.

3.4.4 Структура бази даних PostgreSQL

У базі даних реалізовано такі таблиці (рисунок 3.5):

– dataset – містить інформацію про датасети, поля: id, name, upload_date;

– dataset_data – зберігає як дані датасетів так і результати аналізу тональності тексту, поля: id, real_rating, review_text, dataset_id, predicted_spacy_rating, predicted_textblob_rating, predicted_transformers_rating, predicted_vader_rating, predicted_spacy_rating_text, predicted_textblob_rating_text, predicted_transformers_rating_text, predicted_vader_rating_text;

– dataset_execution_time – зберігає час, затрачений на обробку датасету NLP-бібліотекою, поля: id, execution_time, library_name, dataset_id.

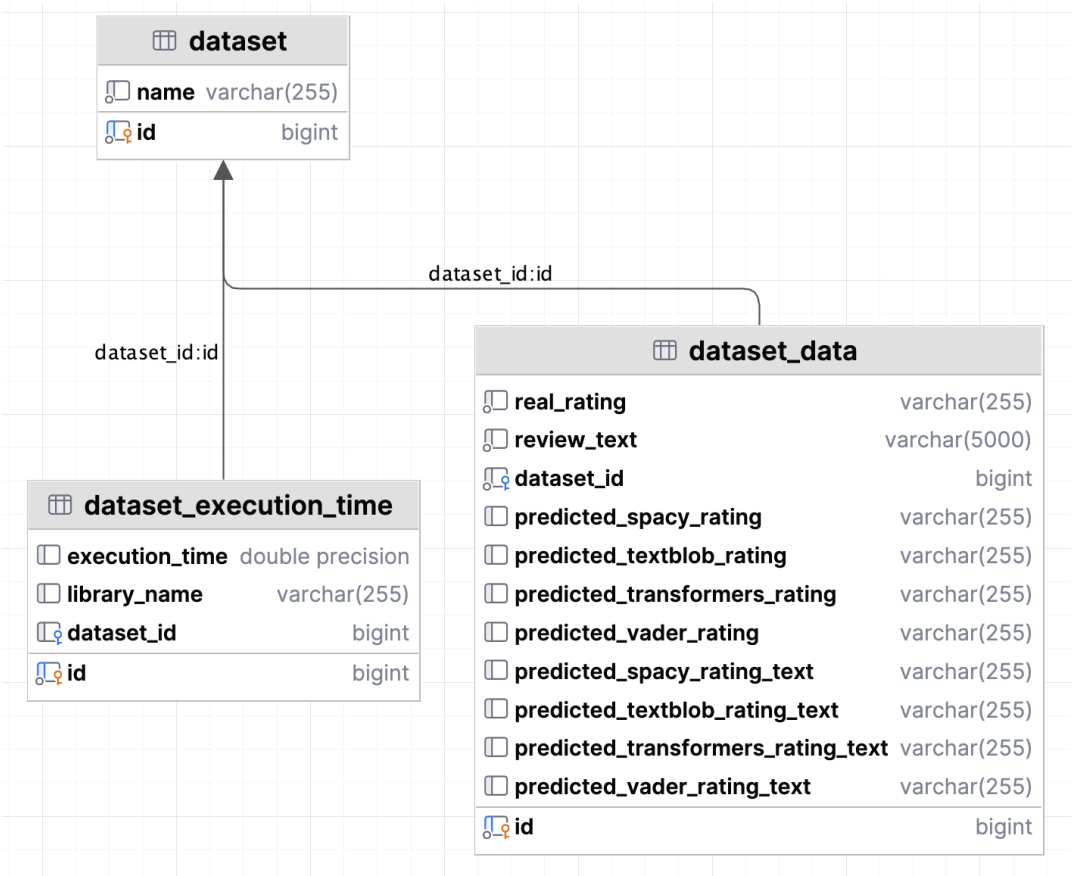

Рисунок 3.5 – Діаграма сутностей і зв’язків

3.5 Реалізація обробки даних за допомогою NLP у Python модулі

Модулі, описані в цьому розділі, інтегрують п’ять популярних бібліотек для аналізу настроїв тексту: VADER, spaCy, TextBlob, Transformers і Stanford NLP. Розроблене програмне забезпечення виконує обробку текстових відгуків і

надає полярні оцінки для кожного відгуку, визначаючи емоційне забарвлення як позитивне, негативне або нейтральне за допомогою цих бібліотек.

Програма розроблена з використанням Flask, що є фреймворком для розробки веб-додатків на Python. Веб-сервіс на Flask відповідає за прийом запитів на обробку даних, їх аналіз і збереження результатів у базу даних PostgreSQL. Програма взаємодіє з п'ятьма різними бібліотеками для NLP, кожна з яких має свої особливості при аналізі відгуків.

VADER (Valence Aware Dictionary and sEntiment Reasoner) бібліотека аналізує текстові відгуки та повертає полярність (позитивний, нейтральний, негативний) на основі лексичних ознак і контексту тексту. Рейтинг полярності обчислюється за допомогою функції SentimentIntensityAnalyzer бібліотеки VADER, яка визначає композитний бал на основі лексичних ознак. Для кожного відгуку результат аналізу зберігається у стовпці predicted_vader_rating у базі даних, де зберігається числове значення композитного балу, та текстова оцінка зберігається у стовпці predicted_vader_rating_text у вигляди positive, neutral або negative значень (рисунок 3.6).

```
def get_vader_rating(review_text):
    scores = vader_analyzer.polarity_scores(review_text)
    compound_score = scores['compound']
    sentiment_text = convert_vader_to_text(compound_score)
    return str(compound_score), sentiment_text
```

Рисунок 3.6 – Код для отримання рейтингу VADER

spaCy бібліотека використовує модель en_core_web_sm для лексичного аналізу тексту. З цією бібліотекою також використано плагін SpacyTextBlob, щоб отримати полярність тексту, яка зберігається як predicted_spacy_rating, та текстова оцінка у predicted_spacy_rating_text колонці у вигляді positive, neutral або negative значень (рисунок 3.7).

```
def get_spacy_rating(review_text):
    doc = nlp(review_text)
    polarity = doc._.polarity
    sentiment_text = "positive" if polarity > 0 else "negative" if polarity < 0 else "neutral"
    return polarity, sentiment_text
```

Рисунок 3.7 – Код для отримання рейтингу spaCy

TextBlob – це ще одна бібліотека для аналізу полярності, яка базується на простому обчисленні значення полярності тексту. Рейтинг тексту визначається за допомогою функції TextBlob, яка аналізує текст на основі своїх вбудованих лексичних характеристик. В результаті кожен відгук отримує полярність та текстову оцінку, яка зберігається в колонці predicted_textblob_rating, а текстова оцінка – у predicted_textblob_rating_text колонці у вигляді positive, neutral або negative значень (рисунок 3.8).

```
def get_textblob_rating(review_text):
    blob = TextBlob(review_text)
    polarity = blob.sentiment.polarity
    sentiment_text = convert_textblob_to_text(polarity)
    return str(polarity), sentiment_text
```

Рисунок 3.8 – Код для отримання рейтингу TextBlob

Бібліотека Transformers від Hugging Face використовує попередньо натреновану модель RoBERTa для аналізу настроїв текстів. Модель працює з класифікацією тексту за категоріями: positive, neutral, negative. Ця бібліотека повертає ймовірність для кожної категорії, яка зберігається у полі predicted_transformers_rating, а текстова оцінка – у predicted_transformers_rating_text колонці у вигляді positive, neutral або negative значень (рисунок 3.9).

```
def get_transformers_rating(review_text):
    result = transformers_pipeline(review_text)
    label = result[0]['label']
    score = result[0]['score']
    if label == 'LABEL_0':
        sentiment_text = 'negative'
    elif label == 'LABEL_1':
        sentiment_text = 'neutral'
    elif label == 'LABEL_2':
        sentiment_text = 'positive'
    return (score, sentiment_text)
```

Рисунок 3.9 – Код для отримання рейтингу Transformers

Stanford NLP використовує модель, натреновану для оцінки настроїв тексту на основі синтаксичного аналізу. Ця бібліотека використовує модель, яка розбиває текст на окремі речення та оцінює їх полярність. Після цього обчислюється середнє значення полярності для всіх речень. Рейтинг за допомогою цієї бібліотеки зберігається в стовпці predicted_stanfordnlp_rating в базі даних, а текстова оцінка – у predicted_stanford_rating_text колонці у вигляді positive, neutral або negative значень (рисунок 3.10).

```
def get_stanfordnlp_rating(review_text):
    doc = stanford_nlp_pipeline(review_text)
    total_sentiment_score = 0
    num_sentences = len(doc.sentences)
    for sentence in doc.sentences:
        total_sentiment_score += sentence.sentiment
    average_sentiment_score = total_sentiment_score / num_sentences
    if average_sentiment_score > 1:
        sentiment_text = 'positive'
    elif average_sentiment_score < 1:
        sentiment_text = 'negative'
    else:
        sentiment_text = 'neutral'
    return average_sentiment_score, sentiment_text
```

Рисунок 3.10 – Код для отримання рейтингу Stanford NLP

Результати аналізу по всім датасетам і бібліотекам зберігаються в базі даних PostgreSQL в таблиці dataset_data, де кожен запис містить текст відгуку та результат аналізу для кожної з бібліотек NLP. Окрім того, для кожної бібліоте-

ки окремо зберігається час виконання обробки у таблиці dataset_execution_time, що дозволяє аналізувати ефективність різних методів. Це дозволяє оптимізувати вибір інструменту для обробки текстових даних, базуючись на часі виконання та точності результатів.

Програма Python модуля приймає запит з інтерфейсу користувача через HTTP POST запит, що містить ідентифікатор набору даних та список NLP бібліотек, які потрібно застосувати для аналізу.

Загалом, цей модуль реалізує ефективне та гнучке рішення для обробки текстових даних за допомогою сучасних методів NLP.

Висновки за розділом 3

У цьому розділі було представлено архітектуру та реалізацію системи аналізу громадської думки за допомогою методів обробки природної мови. Описані модулі та їх взаємодія забезпечують масштабованість і модульність системи, що дозволяє ефективно обробляти великі обсяги даних та інтегрувати нові інструменти аналізу. Реалізація системи з використанням технологій Spring Boot, React, Python, PostgreSQL гарантує зручний користувацький інтерфейс та надійну обробку й збереження даних, що є важливим елементом для подальшої роботи та розвитку системи.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ ТА ЇХ АНАЛІЗ

4.1 Оцінка точності NLP методів за датасетами

На першому етапі експерименту проводилася оцінка точності передбачення для кожної NLP бібліотеки на трьох різних датасетах. Результати точності були виміряні у відсотках, і для кожної бібліотеки було обчислено її точність по кожному датасету. Значення точності наведено на рисунку 4.1 та у таблиці 4.1.

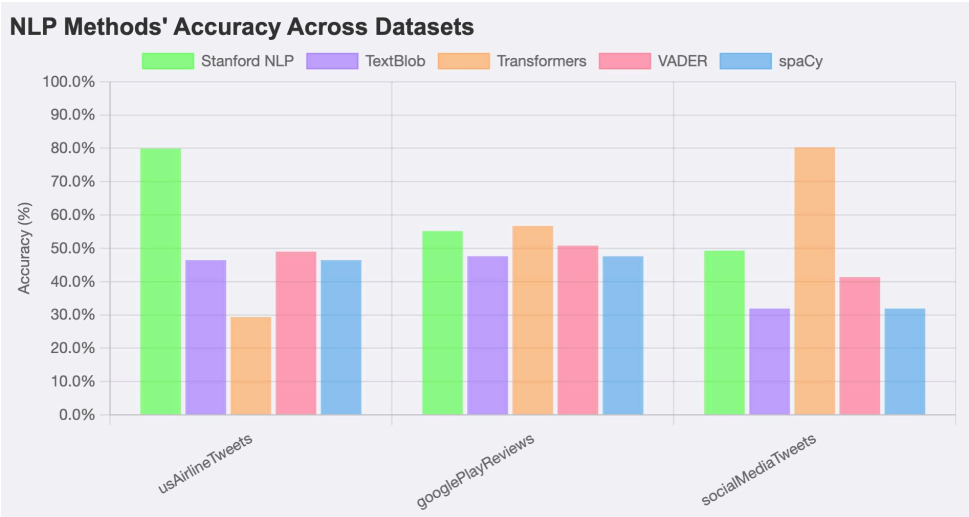

Рисунок 4.1 – Результат розподілу точності прогнозів тональностей тексту NLP-бібліотек для різних датасетів

Таблиця 4.1 – Точність прогнозованих тональностей тексту за допомогою різних NLP-бібліотек на різних датасетах

Датасет	Точність, %				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
US Airline Tweets	79,94	46,42	29,37	48,97	46,42
Google Play Reviews	55,16	47,60	56,71	50,78	47,60
Social Media Tweets	49,30	31,90	80,30	41,33	31,90

Ці результати показують, що кожна з бібліотек має свої сильні і слабкі сторони. Наприклад, Stanford NLP показує відносно високу точність на US Airline Tweets та Social Media Tweets, тоді як Transformers демонструє вищу точність на Social Media Tweets, але низьку на US Airline Tweets.

Наступним кроком було обчислення середньої точності для кожної бібліотеки по всіх датасетах. Це дозволяє отримати загальну картину ефективності кожної бібліотеки в контексті всіх використовуваних даних. Середні результати розподілу точності наведено на рисунку 4.2 та у таблиці 4.2.

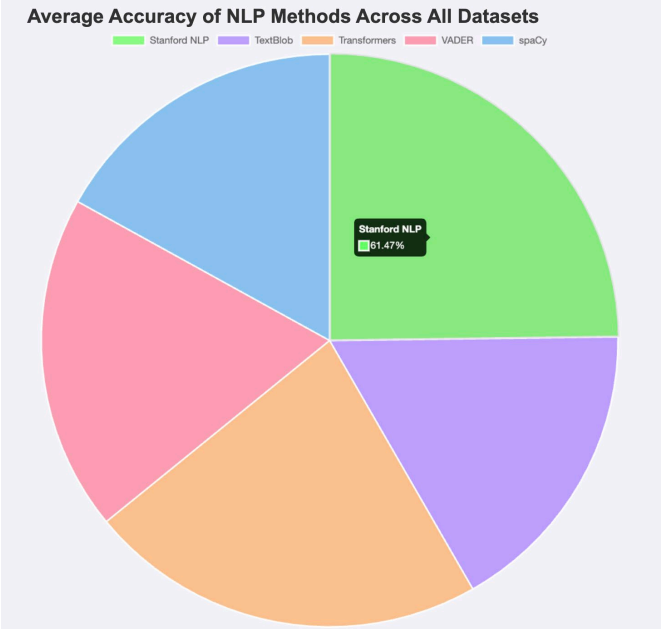

Рисунок 4.2 – Результат середньої точності прогнозів по всім датасетам

Таблиця 4.2 – Середня точність прогнозів по всім датасетам

Датасет	Точність, %				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
Середня точність, %	61,47	41,97	55,46	47,03	41,97

На основі отриманих результатів для всіх датасетів можна виокремити два лідери за точністю – це Stanford NLP та Transformers, які показали середні значення 61,47% та 55,46% відповідно. Ці бібліотеки демонструють найкращі

результати за точністю передбачення, що підтверджується їх високою продуктивністю на таких датасетах, як US Airline Tweets та Social Media Tweets. Це може свідчити про їх здатність краще справлятися з більш складними та варіативними текстами, що особливо важливо при аналізі відгуків і твітів, що містять жаргон або неформальний стиль спілкування.

Однак варто зазначити, що VADER продемонстрував досить швидкі результати, займаючи 47,03% за точністю, що в поєднанні з його високою ефективністю (буде продемонстровано далі) може бути корисним вибором для задач, де швидкість обробки даних важливіша за високу точність. Бібліотеки TextBlob та spaCy показали однакові результати з точністю 41,97%, що може свідчити про їх обмежену здатність до обробки складних синтаксичних конструкцій або текстів з іронічним чи саркастичним контекстом.

4.2 Оцінка часу виконання NLP методів за датасетами

Для більш глибокого аналізу було оцінено не лише точність, але й час виконання обробки кожного датасету різними бібліотеками. Значення витраченого часу наведені на рисунку 4.3 та у таблиці 4.3.

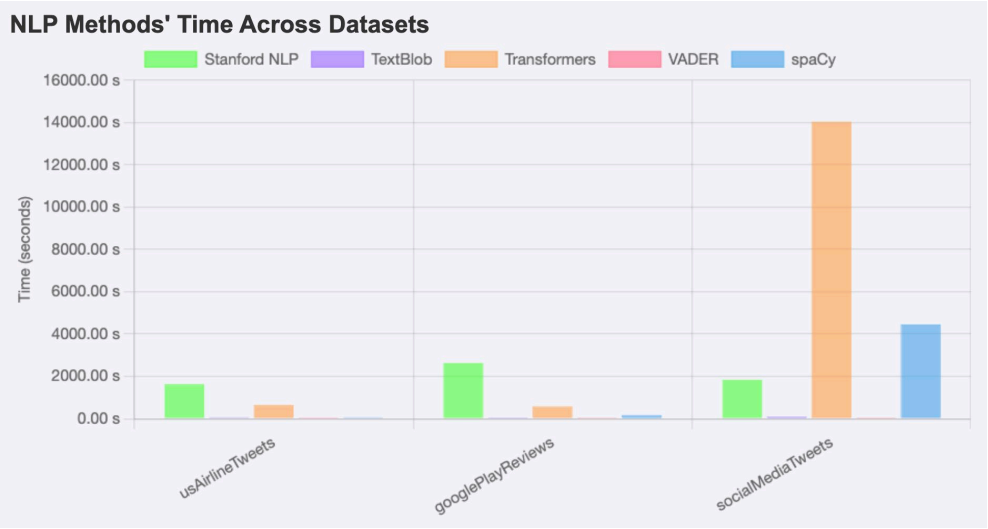

Рисунок 4.3 – Результат розподілу часу виконання для різних датасетів

Таблиця 4.3 – Розподіл часу виконання для різних датасетів

Датасет	Час виконання, s				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
US Airline Tweets	1633,70	72,88	647,03	1,95	63,18
Google Play Reviews	2641,40	58,77	585,31	4,43	176,34
Social Media Tweets	1844,04	98,57	14056,22	2,56	4467,38

Згідно з результатами оцінки часу виконання, найбільший час на обробку був витрачений бібліотекою Transformers при роботі з датасетом Social Media Tweets, де виконання тривало 14056 секунд. Це значно перевищує час виконання інших методів на цьому ж датасеті, таких як spaCy, який обробив дані за 4467 секунд. Інші бібліотеки показали набагато кращі результати по часу: для US Airline Tweets та Google Play Reviews бібліотеки Stanford NLP та TextBlob виконали аналіз значно швидше, витративши відповідно від 63 секунд до 2,5 хвилин на обробку одного датасету.

Ці результати вказують на те, що в залежності від складності даних та розміру датасету деякі методи можуть значно відрізнятись за часом виконання. Бібліотеки, які здатні обробляти великі обсяги даних швидше (як VADER і TextBlob), можуть бути кориснішими для роботи з менш складними або великими наборами даних, тоді як більш складні моделі, такі як Transformers та Stanford NLP, потребують більше часу для обробки складних текстів.

Це також може свідчити про необхідність подальшої оптимізації цих методів для підвищення їх ефективності, зокрема для обробки великих наборів соціальних медіа твітів, де час виконання критично важливий.

Наступним кроком була побудова графіку часу виконання для всіх бібліотек по всім датасетам. Результати сумарного часу виконання наведені на ри-

сунку 4.4 та у таблиці 4.4.

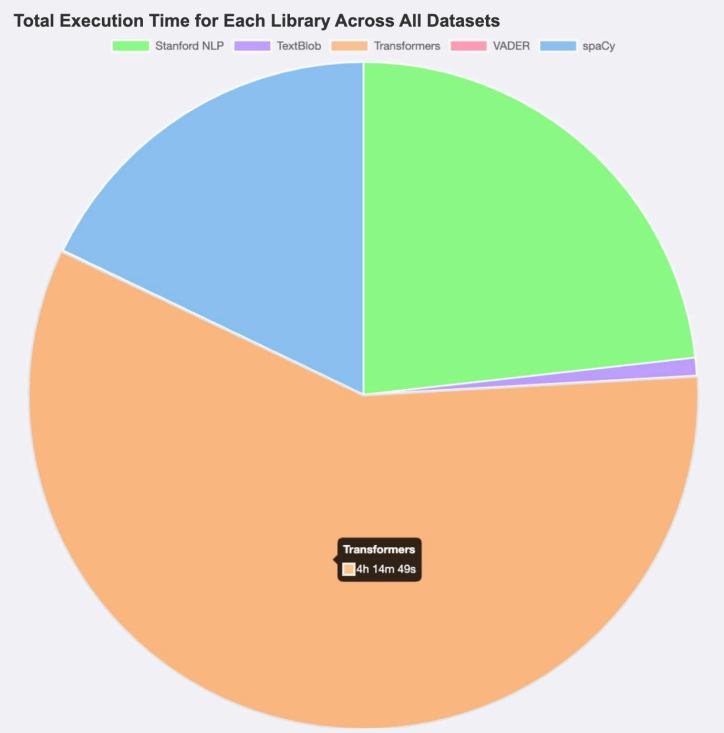

Рисунок 4.4 – Результат розподілу часу виконання по всім датасетах

Таблиця 4.4 – Сумарній час прогнозів по всім датасетам

Датасет	Сумарний час				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
Сумарний час по всім датасетам	1h 41m 59s	3m 50s	4h 14m 49s	9s	1h 18m 27s

Аналіз сумарного часу виконання прогнозів показав значні відмінності у продуктивності NLP-бібліотек. Найшвидшою бібліотекою виявився VADER із сумарним часом виконання 9 секунд, що значно переважає інші бібліотеки. Водночас Transformers демонструє найповільніший час виконання у 4 години 14 хвилин 49 секунд, що може бути наслідком більш складних алгоритмів моделі.

Ці результати підтверджують, що вибір бібліотеки залежить від пріоритетів проєкту: точності або часу виконання. VADER є оптимальним варіантом для швидкого аналізу, тоді як Transformers може бути доцільним для завдань, де

точність є ключовою метою, незважаючи на високі витрати часу.

4.3 Ефективність NLP методів за датасетами

Для порівняння ефективності кожної бібліотеки обчислювалася оцінка ефективності, яка визначалася як відношення точності до часу виконання для кожної бібліотеки на кожному датасету. Значення ефективності NLP-бібліотек по датасетах наведені на рисунку 4.5 та у таблиці 4.5.

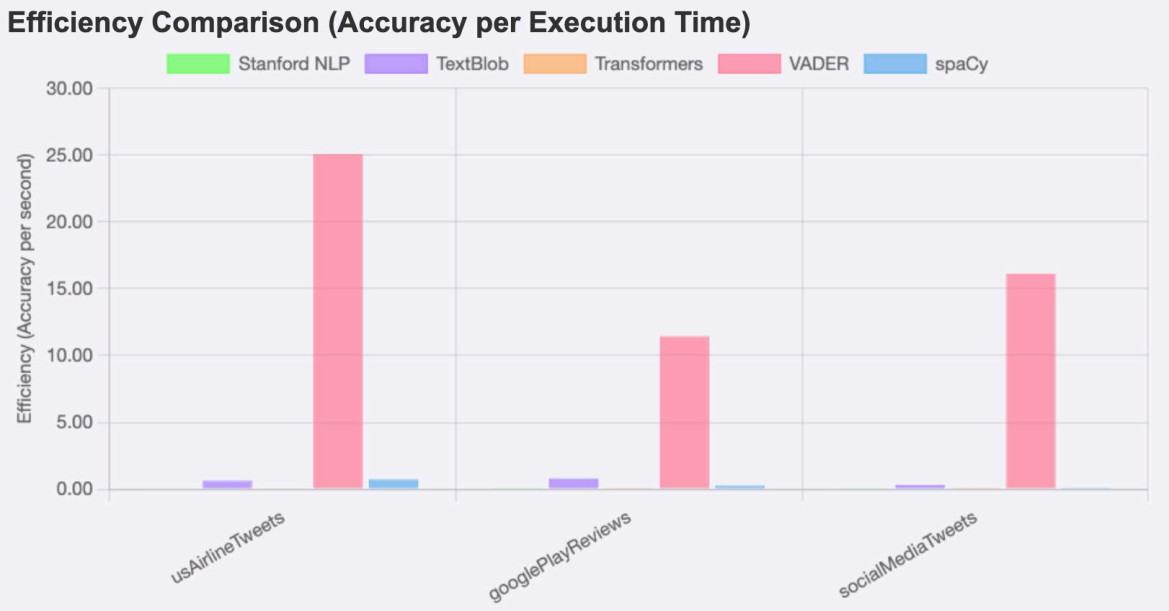

Рисунок 4.5 – Ефективність NLP-бібліотек по датасетах

Таблиця 4.5 – Ефективність NLP-бібліотек по датасетах

Датасет	Ефективність (точність / час виконання)				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
US Airline Tweets	0,049	0,637	0,045	25,112	0,735
Google Play Reviews	0,021	0,81	0,097	11,463	0,27
Social Media Tweets	0,027	0,324	0,006	16,145	0,007

Бібліотека VADER продемонструвала найвищу ефективність серед пред-

ставлених інструментів, особливо на датасеті US Airline Tweets, де її показник досягає 25,112. Це свідчить про високу швидкість роботи бібліотеки разом із задовільною точністю. TextBlob посіла друге місце за ефективністю, показуючи стабільно високі результати на всіх датасетах, зокрема на Google Play Reviews (0,81). У свою чергу, Transformers продемонструвала найменшу ефективність через значний час обробки, а Stanford NLP показала помірні результати.

Таким чином, результати свідчать про значну перевагу VADER у швидкості обробки даних, тоді як TextBlob забезпечує баланс між точністю та продуктивністю. Вибір бібліотеки має залежати від вимог до точності або часу обробки в конкретному сценарії. Значення середньої ефективності NLP-бібліотек по датасетах наведені на рисунку 4.6 та у таблиці 4.6.

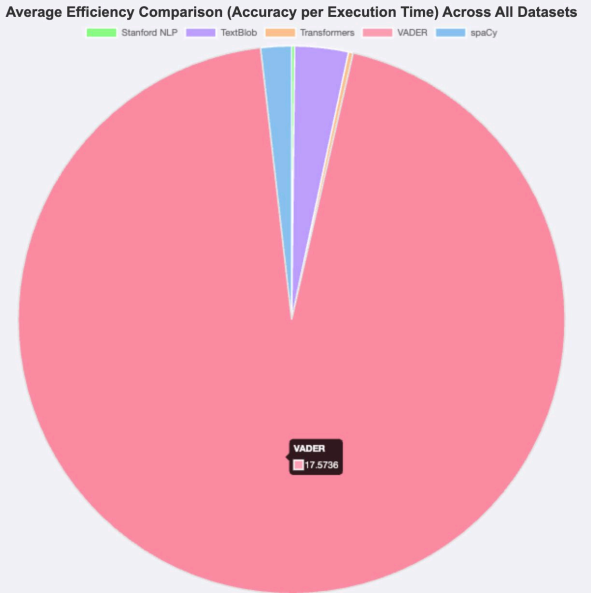

Рисунок 4.6 – Середня ефективність NLP-бібліотек по датасетах

Таблиця 4.6 – Середня ефективність NLP-бібліотек по датасетах

Датасет	Середня ефективність				
	Stanford NLP	TextBlob	Transformers	VADER	spaCy
Середня ефективність	0,0322	0,5902	0,0493	17,5736	0,3373

Середня ефективність різних NLP-бібліотек значно варіюється. Найкра-

щий результат продемонструвала бібліотека VADER, яка досягла показника 17,5736. Це свідчить про її здатність забезпечувати високу продуктивність завдяки оптимальному співвідношенню точності й часу виконання. На другому місці за ефективністю знаходиться spaCy із середнім значенням 0,3373, яка забезпечує збалансовані показники на всіх датасетах.

Найменшу ефективність продемонструвала Stanford NLP, середній показник якої становить 0,0322. Це пов’язано із високими витратами часу на обробку даних. TextBlob та Transformers займають проміжні позиції, показуючи відповідно 0,5902 і 0,0493.

Таким чином, VADER є безперечним лідером серед представлених бібліотек за середньою ефективністю, що робить її найбільш придатною для швидкого аналізу великих обсягів даних. Однак, для задач із вищими вимогами до точності можна розглянути інші інструменти, залежно від потреб конкретного проєкту.

Висновки за розділом 4

Дані дослідження показують, що ефективність різних NLP-бібліотек значно варіюється залежно від методів обробки та умов виконання. Для аналізу використовувалися складні датасети, які містять елементи іронії, сленгу, скорочень та декілька речень, які потребують однієї загальної оцінки. Це ускладнює обробку та вимагає високої точності моделей.

На основі результатів оцінки точності (без урахування часу виконання) Stanford NLP є лідером із середнім показником точності 61,47% серед усіх датасетів. На другому місці знаходиться Transformers, яка досягла середньої точності 55,46%, демонструючи високу якість аналізу. Третю позицію посідає VADER із точністю 47,03%, що є конкурентоспроможним результатом у контексті швидкого аналізу.

Проте, якщо враховувати час виконання, ситуація змінюється. Найпові-

льнішою бібліотекою стала Transformers, яка обробила датасети за 4 години 14 хвилин 49 секунд, тоді як Stanford NLP потребувала 1 годину 41 хвилину 59 секунд. Найшвидшою бібліотекою став VADER, завершуючи аналіз усього за 9 секунд. Це робить його найефективнішим інструментом у співвідношенні точності до часу виконання.

Таким чином, якщо враховувати тільки точність, Stanford NLP і Transformers демонструють найкращі результати, що робить їх ідеальними для аналізу складних текстів. Проте ці моделі потребують значно більше часу та обчислювальних ресурсів. У той же час VADER є лідером за ефективністю, завдяки низьким вимогам до ресурсів і швидкому виконанню, що робить його придатним для задач, де швидкість обробки є критичною.

Варто також зазначити, що для аналізу текстів із такими особливостями, як іронія чи сленг, бібліотеки з вищою точністю (наприклад, Stanford NLP та Transformers) забезпечать більш якісний результат. Однак, для досягнення оптимальних показників їх роботи необхідна додаткова оптимізація та можливе покращення обчислювальної інфраструктури.

ВИСНОВКИ

У кваліфікаційній роботі досліджено ефективність сучасних методів обробки текстових даних на основі NLP-бібліотек для задач аналізу тональності текстів. Проведений аналіз п’яти популярних NLP-бібліотек (Stanford NLP, TextBlob, Transformers, VADER, spaCy) продемонстрував їх високий рівень у сфері аналізу тексту. Stanford NLP досягла найвищої точності (61,47%) серед усіх досліджених бібліотек, що вказує на її актуальність для задач, де якість аналізу є критичною. Transformers показала другий результат за точністю (55,46%) і характеризується значною обчислювальною складністю, адже потребувала найбільшого часу для обробки даних (4 години 14 хвилин 49 секунд). VADER, хоч і поступається за точністю (47,03%), забезпечила найшвидше виконання задачі (9 секунд), що робить її лідером за ефективністю. Дослідження складних текстових датасетів, що включають іронію, сленг, скорочення та багатослівні речення, показало, що Stanford NLP та Transformers є найкращими за якістю обробки таких даних, хоча для цього їм потрібен значно більший час і додаткова оптимізація.

Розроблена тримодульна система забезпечує автоматизацію процесу аналізу тональності текстів, включаючи обробку даних, розподіл задач та візуалізацію результатів. Система має перспективи використання в бізнес-аналітиці, автоматизації аналізу клієнтських відгуків, дослідженнях суспільної думки та інших сферах, де потрібен аналіз текстових даних.

Результати роботи сприяють кращому розумінню ефективності різних NLP-бібліотек, зокрема у контексті точності та часу виконання. Запропоноване рішення дозволяє проводити порівняння бібліотек з огляду на ці параметри, що дає змогу вибирати найбільш оптимальні бібліотеки залежно від вимог до точності та часу обробки даних. Це підвищує ефективність використання ресурсів у задачах аналізу тексту. Система може бути корисною для наукових досліджень, а також має соціально-економічну значущість, оскільки її можна використовувати для оптимізації процесів у аналітичних та інших пов’язаних сферах.

Подальші дослідження можуть бути спрямовані на оптимізацію роботи NLP-бібліотек для задач аналізу багатомовних текстів, інтеграцію додаткових інструментів для аналізу контексту та стилістики тексту, а також розширення системи для підтримки нових форматів даних і вдосконалення користувацького інтерфейсу. Таким чином, результати роботи демонструють високу ефективність і практичну цінність запропонованого підходу до аналізу текстових даних, що має значний потенціал для впровадження у різних сферах діяльності.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ziukin D. Analysis of public opinion using natural language processing (NLP) methods. *III International Scientific and Practical Conference «LEARNING & TEACHING in the World after the War»*, Kharkiv, Ukraine, November 8, 2024. P. 251.

2. Катренко А. В. Системний аналіз. Львів : “Новий світ – 2000”, 2011. 396 с.

3. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. London : Pearson, 2023. 1024 p.

4. Goldberg Y. Neural Network Methods for Natural Language Processing. San Rafael, CA : Morgan & Claypool, 2017. 309 p.

5. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. 2nd ed. Sebastopol, CA : O’Reilly Media, 2019. 512 p.

6. Twitter US Airline Sentiment. URL: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment (дата звернення: 25.11.2024).

7. Sentiment Analysis dataset-Google Play App Reviews. URL: https://www.kaggle.com/datasets/farhaouimouhamed/sentiment-analysis-datasetgoogle-play-app-reviews (дата звернення: 25.11.2024).

8. training.1600000.processed.noemoticon.csv. URL: https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv (дата звернення: 25.11.2024).

9. Уолс К. Spring in Action. 5th edition. New York : Manning Publications, 2018. 520 p.